Exact Bayesian inference for level-set Cox processes with piecewise constant intensity function

Flávio Gonçalves

joint work with Bárbara Dias (UFRRJ)

UFMG

COLMEA

16/11/2022

Content

1. Introduction
2. Background methodologies
3. Spatial Model
4. Bayesian inference
5. Simulations and Applications
6. Conclusions
7. References

Content

1. Introduction
2. Background methodologies
3. Spatial Model
4. Bayesian inference
5. Simulations and Applications
6. Conclusions
7. References

Motivation

- Statistical models for point pattern data are widely used in a variety of areas.
- Most popular model: Poisson process (PP). Important subclass: Cox processes.
  Our approach: Cox processes with piecewise constant IF with flexible space partition.
- Efficient inference methodologies have been proposed for the unidimensional case using continuous time Markov chains to model the IF.
- Existing methodologies for the multidimensional case still rely on discrete approximations leading to systematic bias and potential model decharacterisation.
- Model: Level-set spatiotemporal Cox process.
  Main contribution: methodology to perform exact Bayesian inference - no discrete approximation is used and Monte Carlo error is the only source of inaccuracy.

Motivational examples

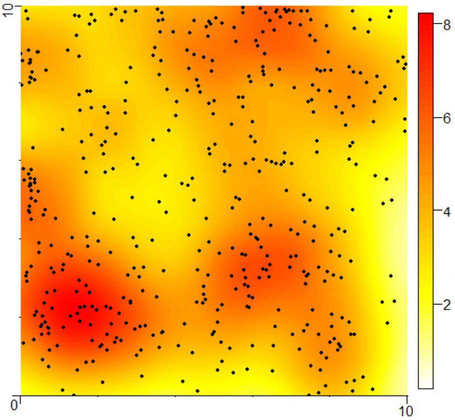


Figure: White oaks in Lansing Woods, USA. Estimated IF via kernel smoothing.

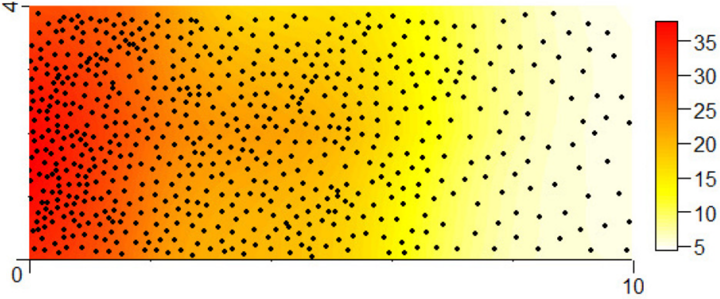


Figure: Particles in a bronze filter section profile. Estimated IF via kernel smoothing.

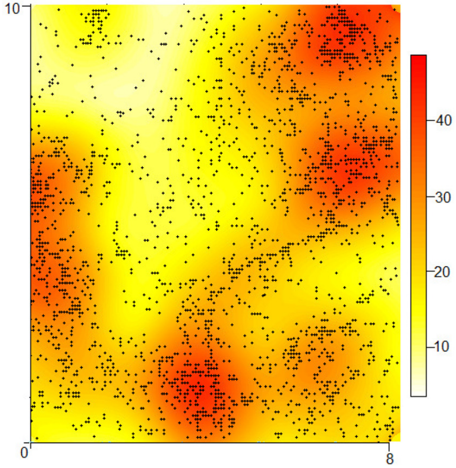


Figure: Fires in a region of New Brunswick, Canada. Estimated IF via kernel smoothing.

Literature review

- Heikkinen and Arjas [1998] and Møller and Rasmussen [2015] use Voronoi tessellation to specify a piecewise constant and Kernel-based structure for the IF, respectively.
- Myllymäki and Penttinen [2010] propose the level-set Cox process with 2 levels.
- Hildeman et al. [2018] generalises the model for more levels and non-constant IF.
- Level set models define a partition of some compact region (in $\mathbb{R}^2$) by means of the levels of a latent Gaussian process.
- Because of the difficulties to perform inference due to the intractability of the actual (infinite-dimensional) model, the two aforementioned papers consider a discrete version of this.
- A regular lattice that models the number of points in each cell as a Poisson distribution. The latent GP is replaced by a multivariate normal with one coordinate per cell.
- "The information on the fine scale behavior of the point pattern is lost".

Our aims

- Exact methodology to perform Bayesian inference for level-set Cox process models in which the IF is piecewise constant [Gonçalves and Dias, 2022].

- Difficulties:
  1. intractability of the likelihood function of the proposed model;
  2. infinite dimensionality of the model's parameter space due to the latent GP.

- Solution: pseudo-marginal MCMC with retrospective sampling.

- Dealing with high computational cost associated to GPs: nearest neighbor Gaussian process (NNGP) [Datta et al., 2016]. Key property: defines a valid GP measure - Bayesian paradigm is preserved.

- This is, to the best of our knowledge, the first work to consider a latent NNGP within a complicated likelihood structure that does not allow for directly sampling from the posterior or full conditional distribution of the NNGP component.

Content

1 Introduction

2 Background methodologies

3 Spatial Model

4 Bayesian inference

5 Simulations and Applications

6 Conclusions

7 References

Pseudo Marginal Metropolis-Hastings

- Suppose that the likelihood is intractable and cannot be evaluated pointwise.
- Andrieu and Roberts [2009]: replace the likelihood by an a.s. positive and unbiased estimator of this in the expression of the a.p. of a MH algorithm - preserves the posterior as the marginal invariant distribution of the chain (integrating out w.r.t. the extra r.v.).
- Define $U \sim q_U$ and $\hat{L}$ such that $E[\hat{L}(y,\theta,U)] = L(\theta,y)$ and $\hat{L} \overset{\text{a.s.}}{>} 0$, $\forall\theta \in \Theta$, $\forall y \in \mathcal{Y}$.

Algorithm 1 PSEUDO MARGINAL METROPOLIS-HASTINGS

1 Propose $\theta' \sim q(.|\theta)$ and $U' \sim q_U$;

2 Accept w.p. $\alpha(\theta,U;\theta',U') = 1 \wedge \dfrac{\hat{L}(y,\theta',U')\pi(\theta')q(\theta|\theta')}{\hat{L}(y,\theta,U)\pi(\theta)q(\theta'|\theta)}$.

Retrospective sampling and infinite-dimensional MCMC

- Retrospective sampling is a simulation technique that changes the natural order of steps to make the algorithm more efficient or even feasible. It is particularly useful to simulate infinite-dimensional r.v.'s.

- The idea is to be able to perform the algorithm (typically accept-reject type) by unveiling only a finite-dimensional representation of the r.v. of interest and to have an efficient recovery algorithm to simulate the remainder of the r.v.

- In our context, we propose an infinite-dimensional retrospective MCMC algorithm. The GP component is sampled retrospectively via PMMH.

Content

1. Introduction
2. Background methodologies
3. Spatial Model
4. Bayesian inference
5. Simulations and Applications
6. Conclusions
7. References

Proposed model

$$
\begin{aligned}
(Y|\lambda_S) &\sim PP(\lambda_S), \\
\lambda(s) &= \sum_{k=1}^{K} \lambda_k I_k(s), \\
S_k &= \{s \in S : c_{k-1} < \beta(s) < c_k\}, \forall k \\
\beta &\sim GP(\mu, \Sigma), \\
\pi(c) &= \mathbb{1}(c_1 < \ldots < c_{K-1}), \\
\lambda &\sim prior
\end{aligned}
$$

- β, c and λ's are assumed to be independent *a priori*.
- Other option: $\lambda(s) = \sum_{k=1}^{K} \kappa(s)\lambda_k I_k(s)$, where $\kappa(s)$ is an offset term.

- The likelihood function of the level-set Cox process model is not identifiable. For each point in the (infinite-dimensional) parameter space, there is an uncountable number of other points that return the same likelihood value.
- This is caused by the non-identification of the scale of the GP. Write $\beta = \mu + \sigma\beta^*$, where $\beta^* \sim N(0, \Sigma(1, \tau^2))$. Any $\mu^* = a\mu + b$, $\sigma^* = a\sigma$ and $c_k^* = b + ac_k$, $b \in \mathbb{R}$, $a \in \mathbb{R}^+$, $\forall k$, defines the same partition and, consequently, the same likelihood.
- Solution: fix either c or (μ, σ^2). We shall adopt the latter.
- Label-switching of the coordinates of λ is unlikely, given the complexity of the sample space.
- The number of levels is fixed based on prior information, the type of structure the researcher expects, or even an empirical analysis of the data.
  Trade-off: model fitting and parsimony.
- The piecewise constant structure allows for a cluster analysis perspective.

NNGP prior for β

- The computational bottleneck of the methodology is sampling the GP. Cost to simulate a d-dimensional normal is $\mathcal{O}(d^3)$.
- Solution: NNGP. Exact in the sense of defining a valid probability measure and, therefore, preserving the Bayesian paradigm.
- Originally designed to approximate a parent GP in classical geostatistical problems in which the (discretely) observed process is either the GP itself or the GP + i.i.d. noise.
- In our context, the GP is latent in a more complex way. But it only determines the partition and not the actual values of the IF. It is reasonable to see the NGPP simply as the GP prior for β and not an approximation for some desirable traditional GP.
- The NNGP is devised from a parent $GP(\mu, \Sigma(\sigma^2, \tau^2))$ by imposing some conditional independence structure that leads to a sparsity.

For a reference set $\mathcal{S} = \{\mathfrak{s}_1, \ldots, \mathfrak{s}_r\}$ and a maximum number m of neighbors,

$$\begin{aligned}
\pi(\beta) &= \pi(\beta_{\mathcal{S}})\pi(\beta_{S\setminus\mathcal{S}}|\beta_{\mathcal{S}}), \\
\pi(\beta_{\mathcal{S}}) &= \pi_{GP}(\beta_{\mathfrak{s}_1})\pi_{GP}(\beta_{\mathfrak{s}_2}|\beta_{\mathfrak{s}_1})\pi_{GP}(\beta_{\mathfrak{s}_3}|\beta_{\mathfrak{s}_1},\beta_{\mathfrak{s}_2})\ldots\pi_{GP}(\beta_{\mathfrak{s}_{m+1}}|\beta_{\mathfrak{s}_1},\ldots,\beta_{\mathfrak{s}_m}) \\
&\quad \pi_{GP}(\beta_{\mathfrak{s}_{m+2}}|\beta_{\mathcal{N}(\mathfrak{s}_{m+2})})\ldots\pi_{GP}(\beta_{\mathfrak{s}_r}|\beta_{\mathcal{N}(\mathfrak{s}_r)}), \\
\pi_{GP}(\beta_{S_0}|\beta_{\mathcal{S}}) &= \prod_{i=1}^{I}\pi_{GP}(\beta s_i|\beta_{\mathcal{N}(s_i)}), \text{ for any finite set } S_0 = \{s_1,\ldots,s_I\} \subset S\setminus\mathcal{S},
\end{aligned}$$

where $\mathcal{N}(\mathfrak{s}_i)$ is the set of the m closest neighbors of $\mathfrak{s}_i$ in $\{\mathfrak{s}_1, \ldots, \mathfrak{s}_{i-1}\}$, for $i \geq m+2$, and $\mathcal{N}(s_i)$ is the set of the m closest neighbors of s_i in $\mathcal{S}$.

- In traditional geostatistical models the reference set is conveniently defined to be the locations of the observations. Not reasonable in our case. We set $\mathcal{S}$ to be a regular lattice on S with $r = 2500$ and $m = 16$.

- The conditional independence among the locations in S_0 allows the parallelisation of the algorithm to sample from this. Our MCMC needs to sample from the NNGP prior in a large set S_0 on every iteration of the algorithm.

Covariance function

- The covariance function $\Sigma(\sigma^2, \tau^2)$ plays an important role in the methodology. We use the powered exponential with exponent $\gamma = 1.95$.

$$Cov(\beta(s), \beta(s')) = \exp\left\{-\frac{1}{2\tau^2}|s - s'|^\gamma\right\}.$$

- The Poisson process likelihood is ill-posed. It increases indefinitely as the IF increases in (infinitesimal) balls centred around the observations and approaches zero outside them.
- The Cox process formulation is a way to regularise the likelihood function by assigning a prior to the IF.
- This prior has great impact on the posterior. The posterior of β is absolutely continuous w.r.t. its prior.
- The likelihood favors the pattern described above which, in turn, favors smaller values of τ^2 (less smooth). So, fixing τ^2 is a reasonable strategy.
- This determines the smoothness of the IF. Typically, partitions with very small regions should be avoided.

Prior on λ

- The prior information GP may not be enough to avoid model identifiability problems. A reasonable solution is to add coherent prior information through the prior of λ.
- Model parsimony: fit models with fewer levels and clearly distinct rates. This is favored by adopting a repulsive prior for λ.
- Prior based on the Rep distribution proposed in Quinlan et al. [2021]. We penalise a scaled version of the differences between the λ_k's.

$$\begin{aligned}
\pi(\lambda) &\propto \left[\prod_{i=1}^{K} \pi_G(\lambda_k)\right] R(\lambda;\rho,\nu), \\
\pi_G(\lambda_k) &\propto \lambda_k^{\alpha_k - 1} e^{-eta_k \lambda_k},\ \alpha_k > 0,\ \eta_k > 0,\ k = 1,\ldots,K, \\
R(\lambda;\rho,\nu) &= \prod_{1 \le k_1 < k_2 \le K} \left(1 - \exp\left\{-\rho\left(\frac{|\lambda_{k_1} - \lambda_{k_2}|}{\sqrt{\lambda_{k_1} + \lambda_{k_2}}}\right)^{\nu}\right\}\right).
\end{aligned}$$

- Repulsive gamma prior - $RG(\alpha,\eta,\rho,\nu)$. Suggestion: $\rho \in [1,5]$ and $\nu = 3$.
- The RG prior is proper and can be useful to identify K.

Content

1. Introduction
2. Background methodologies
3. Spatial Model
4. Bayesian inference
5. Simulations and Applications
6. Conclusions
7. References

Bayesian inference

Likelihood function and posterior density:

$$L(\theta, Y) \propto \exp\left\{-\sum_{k=1}^{K} \lambda_k \mu_k\right\} \prod_{k=1}^{K} (\lambda_k)^{|Y_k|},$$

μ_k and $|Y_k|$ are the area and number of observations in region S_k.

$$\pi(\theta, Y) \propto \exp\left\{-\sum_{k=1}^{K} \lambda_k \mu_k\right\} \left[\prod_{k=1}^{K} (\lambda_k)^{|Y_k|} \pi(\lambda_k)\right] \left[\prod_{k=1}^{K-1} \pi(c_k)\right] \pi_{GP}(\beta).$$

Intractability of $\pi(\theta, Y)$:

- The Gaussian process β is infinite-dimensional. Solution: retrospective sampling.
- The density $\pi_{GP}(\beta)$ is intractable. Solution: use as proposal distribution that cancels out with the prior density in the expression of the acceptance probability.
- μ_k is intractable. Solution: pseudo-marginal with unbiased estimation of the likelihood - $M = \exp\left\{-\sum_{k=1}^{K} \lambda_k \mu_k\right\}$ - via *Poisson Estimator*.
- unbiased estimators for the μ_k's can be easily obtained using uniform r.v.'s on S, for M nonetheless...
- The pseudo-marginal estimator ought to be devised in a way that the auxiliary r.v. has a θ-free distribution so that we can block the algorithm in a Gibbs sampling [Murray and Graham, 2016].

Poisson Estimator

Proposition 1

Define $N^* \sim PP(1)$ *in the cylinder with base* S *and height in* $[0, +\infty)$ *and let* $N = g(N^*, \lambda^*)$ *be the projection on* S *of the points from* N^* *that are below* $\lambda^* = (\delta\lambda_M - \lambda_m)$*, for* $\lambda_M = \max_k\{\lambda_k\}$ *and* $\lambda_m = \min_k\{\lambda_k\}$*. Then, for any* $\delta > 1$*, an unbiased and a.s. positive estimator for* M *is given by*

$$\hat{M} = e^{-\mu(S)\lambda_m} \prod_{k=1}^{K} \left(\frac{\delta\lambda_M - \lambda_k}{\delta\lambda_M - \lambda_m} \right)^{|N_k|},$$

where $\mu(S)$ *is the area of* S *and* $|N_k|$ *is the number of points from* N *in* S_k*.*

Proposition 2

Estimator $\hat{M}$ *has a finite variance which is a decreasing function of* δ*.*

$$
\begin{aligned}
E_{|N|,I}[\hat{M}] &= E_{|N|,I}\left[e^{-\mu(S)\lambda_m}\prod_{k=1}^{K}\left(\frac{\delta\lambda_M-\lambda_k}{\delta\lambda_M-\lambda_m}\right)^{|N_k|}\right] \\
&= E_{|N|,I}\left[e^{-\mu(S)\lambda_m}\prod_{n=1}^{|N|}\left(\frac{\delta\lambda_M-\sum_{k=1}^{K} I_{nk}\lambda_k}{\delta\lambda_M-\lambda_m}\right)\right] \\
&= e^{-\mu(S)\lambda_m}E_{|N|}\left[\left(\frac{\mu(S)\delta\lambda_M-\sum_{k=1}^{K}\mu_k\lambda_k}{\mu(S)(\delta\lambda_M-\lambda_m)}\right)^{|N|}\right] \\
&= e^{-\mu(S)(\lambda_m+\delta\lambda_M-\lambda_m)}\sum_{j=0}^{\infty}\frac{\left(\mu(S)\delta\lambda_M-\sum_{k=1}^{K}\mu_k\lambda_k\right)^{j}}{j!} \\
&= e^{-\sum_{k=1}^{K}\mu_k\lambda_k} = M.
\end{aligned}
$$

- In a retrospective sampling context, it is N that determines the locations at which β is to be simulated, besides the locations from Y (and $\mathcal{S}$).

- The mean number of locations from N is $(\delta\lambda_M - \lambda_m)\mu(S)$.

- Trade-off in the choice of δ: if it increases, the variance of $\hat{M}$ decreases (improves the mixing of the PSMH chain - in principle) but the computational cost per iteration increases.

- An increase in δ also increases the (expected) dimension of N, which may have a negative impact in the mixing of the MCMC, specially in a Gibbs sampling that samples N and β separately.

Conceptual and practical pseudo-marginal MCMCs

1. Propose a move $(\theta, N^*) \to (\ddot{\theta}, \ddot{N}^*)$ from a density $q(\ddot{\theta}, \ddot{N}^*|\theta, N^*) = q(\ddot{\theta}|\theta)q(\ddot{N}^*)$, where $q(\ddot{N}^*) \sim PP(1)$.

2. Accept a move with probability

$$1 \wedge \left(\frac{\hat{\pi}(\ddot{\theta}; \ddot{N}^*)}{\hat{\pi}(\theta; N^*)} \frac{q(\theta|\ddot{\theta})}{q(\ddot{\theta}|\theta)} \right).$$

- Bound to be inefficient given the complexity of the coordinates. Simple solution though.
- Block the coordinates - Gibbs sampling with PMMH steps - same a.p.
- N^* can be a block because its distribution does not depend on θ. N^* in infinite, but we only need N to compute the a.p.
- Blocks*: N^*, β, λ, c, with retrospective sampling for β and N^*.

Sampling N^*

- Sampling N^* from $q(N^*)$ is bound to lead to low acceptance rate.

- Update N^* below and above λ^*, separately. Latter: sampled retrospectively (if needed) from $q(N^*)$ w.p. 1.

- Former: split S into L (regular) cells and update N^* in each cylinder separately. Under $q(N^*)$, N^* (N) is independent among the L cylinders and follows a $PP(1)$ ($PP(\lambda^*)$) in each of them.

- Optimal scaling problem w.r.t. L. Empirical analyses suggest L so that the average a.r. is around 0.8.

Sampling β

- Retrospective sampling: sampled at a finite collection of locations which are enough to perform all the steps of the MCMC algorithm - $\mathcal{S}$, Y and N.

- 1. Impossible to sample β directly from its full conditional. 2. the proposal has to imply in a tractable expression for the a.p. - requires term $\pi_{GP}(\beta)$ to be canceled out.

- The conditional independence structure of the NNGP demands extra care to specify this proposal. An independent proposal ($\pi_{GP}(\beta)$) would be inefficient.

- The preconditioned Crank–Nicolson (pCN) proposal [Cotter et al., 2013]:

$$\begin{aligned} \ddot{\beta}(s) &= \sqrt{1-\varsigma^2}\beta(s) + \varsigma\varepsilon(s), \quad s \in S, \\ \varepsilon &\sim NNGP(0, \tilde{\Sigma}). \end{aligned} \tag{1}$$

- In a finite-dimensional context, the pCN proposal differs slightly from the traditional centred random walk, but cancels out with the prior MN density.
  The pCN proposal is valid in the infinite-dimensional context whereas the centred random walk is not. ς^2 is tuned to get a.r. approx. 0.234.

Sampling λ and c

- The proposal for λ is a Gaussian random walk with a properly tuned covariance matrix, based on the respective empirical covariance matrix of the chain, to have the desired acceptance rate - varying from 0.4 to 0.234 according to the dimension of λ.

- Parameter c is jointly sampled from a uniform random walk proposal with a common (and properly tuned) length for each of its components.

Important computational aspects

- Despite the NNGP prior, the computational cost may still be compromised by a large accumulation of points from β resulting from the simulation of extra points to update λ and N^* and successive rejections of β.

- Solution: virtual update steps to update β in $S \setminus \{\mathcal{S}, Y, N\}$ (prior proposal and a.p. 1). In practice, simply delete all the values of β at $S \setminus \{\mathcal{S}, Y, N\}$. This strategy also allows us to retrospectively sample β from its GP prior, instead of the pCN proposal (which would be impractical), on the update steps of λ and N^*. A virtual update is performed every time $S \setminus \{\mathcal{S}, Y, N\}$ is non-empty after an update step.

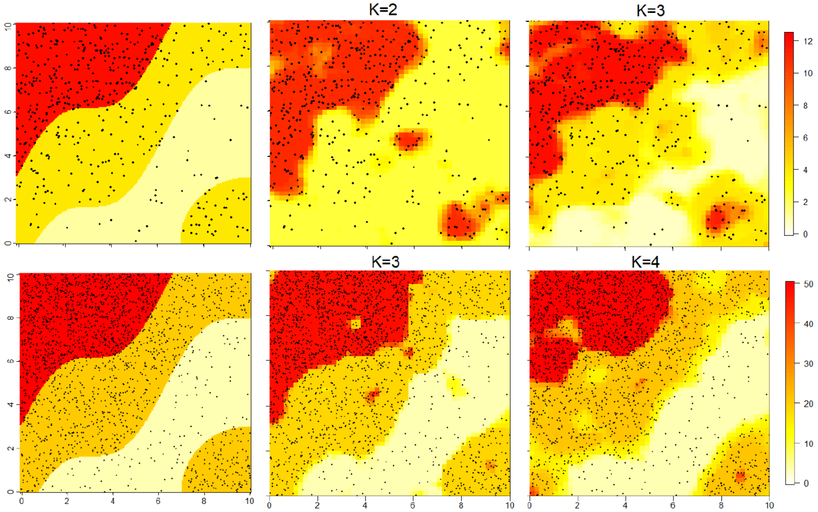
- Choosing δ: mean number of points from N under the pseudo-marginal distribution. Suggestion: ≈ 6000.

- The step to update N is parallelised among the L cells. Sampling β in $S \setminus \mathcal{S}$ is also parallelised.

Content

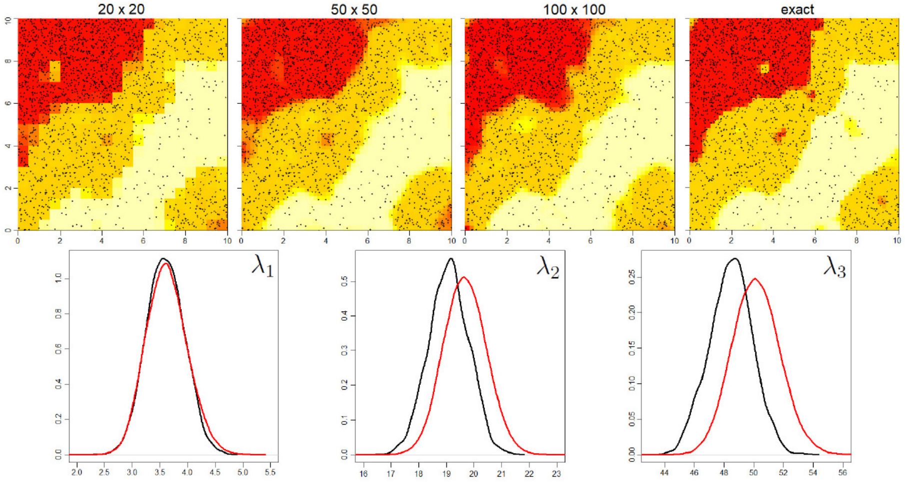
1. Introduction
2. Background methodologies
3. Spatial Model
4. Bayesian inference
5. Simulations and Applications
6. Conclusions
7. References

Simulated examples

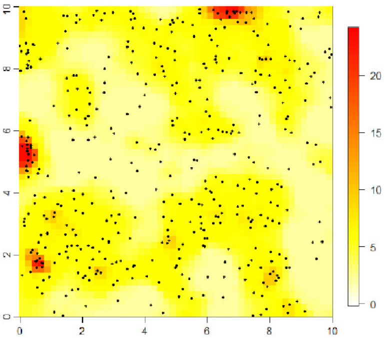


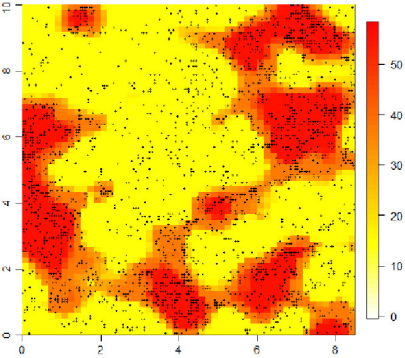
	Example 1		Example 2	
	$K = 2$	$K = 3$	$K = 3$	$K = 4$
λ_1	2.17(0.20)	0.67(0.18)	3.60(0.34)	3.37(0.39)
λ_2	10.84(0.65)	3.99(0.31)	19.09(0.72)	13.76(1.18)
λ_3		11.97(0.74)	48.45(1.44)	21.45(0.84)
λ_4				50.05(1.45)

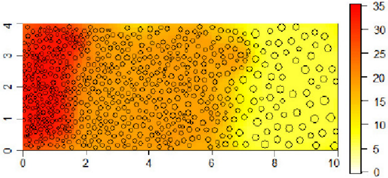
Comparison to discrete method



Applications







	White oak	Bronze filter	NB fires
λ_1	22.48(4.63)	33.27(2.86)	55.11(1.95)
λ_2	6.07(0.42)	18.62(1.15)	37.45(1.72)
λ_3	1.97(0.25)	6.47(0.79)	13.40(0.53)

Content

1 Introduction

2 Background methodologies

3 Spatial Model

4 Bayesian inference

5 Simulations and Applications

6 Conclusions

7 References

Final remarks

- Novel methodology to perform exact Bayesian inference for level-set Cox processes with piecewise constant IF - flexible model and exact inference.
- Infinite-dimensional pseudo-marginal MCMC algorithm with retrospective sampling. Efficient proposal distribution for the latent GP. Computational cost issues dealt by a NNGP and virtual update steps.
- A variety of issues related to the efficiency of the proposed MCMC algorithm are discussed and empirically explored through simulations.
- Spatiotemporal extension - temporal dependency on the GP (β) and on the levels (λ).
- Directions for future work: more complex covariance structures such as non-stationarity; LSCP with non-constant IF [Gonçalves and Gamerman, 2018].

References I

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, pages 697–725, 2009.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, 28:424–446, 2013.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812, 2016.

F. B. Gonçalves and B. C. C. Dias. Exact bayesian inference for level-set cox processes with piecewise constant intensity function. *To appear in Journal of Computational and Graphical Statistics*, 2022.

F. B. Gonçalves and D. Gamerman. Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes. *Journal of the Royal Statistical Society, Series B*, 80:157–175, 2018.

J. Heikkinen and E. Arjas. Non-parametric bayesian estimation of a spatial poisson intensity. *Scandinavian Journal of Statistics*, 25(3):435–450, 1998.

A. Hildeman, D. Bolin, J. Wallin, and J. B. Illian. Level set cox processes. *Spatial Statistics*, 28: 169–193, 2018.

J. Møller and J. G. Rasmussen. Spatial cluster point processes related to poisson–voronoi tessellations. *Stochastic environmental research and risk assessment*, 29(2):431–441, 2015.

References II

Iain Murray and Matthew Graham. Pseudo-marginal slice sampling. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *JMLR: W&CP*, pages 911–919, Cadiz, Spain, 2016.

M. Myllymäki and A. Penttinen. Bayesian inference for gaussian excursion set generated cox processes with set-marking. *Statistics and Computing*, 20:305–315, 2010.

J. J. Quinlan, F. A. Quintana, and G. L. Page. On a class of repulsive mixture models. *Test*, 30: 445–461, 2021.

Obrigado!

fbgoncalves@est.ufmg.br

www.est.ufmg.br/∼fbgoncalves/eng